Voice-QA: Evaluating the Impact of Misrecognized Words on Passage Retrieval

Marcos Calvo, Davide Buscaldi, Paolo Rosso

Voice-QA: Evaluating the Impact of Misrecognized Words on Passage Retrieval

Marcos Calvo
DSIC
Universitat Politècnica de València
Camí de Vera s/n
Valencia, Spain
mcalvo@dsic.upv.es

Davide Buscaldi
IRIT
Université Paul Sabatier
118 rue de Narbonne,
F-31062
Toulouse, France
davide.buscaldi@irit.fr

Paolo Rosso
DSIC
Universitat Politècnica de València
Camí de Vera s/n
Valencia, Spain
prosso@dsic.upv.es

ABSTRACT

Question Answering is an Information Retrieval task where the query is formulated as a natural language question and the expected result is a concise answer. Voice-activated Question Answering systems represent an interesting application, where the input question is formulated by speech. In these systems, an Automatic Speech Recognition module is used to recognize question and transform it in a written form. Because of this process, recognition errors can be introduced, producing a significant effect on the answer retrieval process. In this work we studied the relationship between some characteristics of misrecognized words and the retrieval results. The characteristics we took into account are the redundancy of a word in the result set and its inverse document frequency calculated over the collection. The results show that the redundancy of a word in the result set may be an important clue on whether an error over that word would produce a deterioration of the retrieval results, at least if a closed entity model is used for speech recognition.

Categories and Subject Descriptors

H.3 [**Information storage and retrieval**]: Information Search and Retrieval; I.2 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, experimentation, performance

1. INTRODUCTION

Question Answering (QA) is an Information Retrieval (IR) task in which the query is posed in natural language and the expected result is a concise answer instead of a list of document. Currently, most QA systems accept written sentences as their input, but in the last years there has been a growing interest in systems where query are formulated by voice [1, 4]. In fact, due to the interest on this kind of applications, some Evaluation Conferences, such as CLEF (Cross-Language Evaluation Forum) competition have included a voice-activated Question Answering task in different languages [6].

A Question Answering system is composed by several modules, corresponding to different steps in the analysis of the question and the search for the answer. In general, a QA system is composed by an analysis module, which determines the type of the question, a Passage Retrieval (PR) module, which uses standard IR techniques to retrieve passages where the answer could be contained, and an answer extraction module, which uses NLP techniques or patterns to extract the answer from the passages. In addition to these modules, if the input of our system are utterances, an Automatic Speech Recognition (ASR) module is also needed, in order to transform the vocal input in a written question. One option is to "plug" the ASR before the QA modules, in such a way that the input to the QA system is the sentence (or the n-best sentences) recognized by the ASR. In Figure 1 we show the architecture of such a system, where the output is given back to the user by means of a Text-To-Speech synthesizer (TTS).

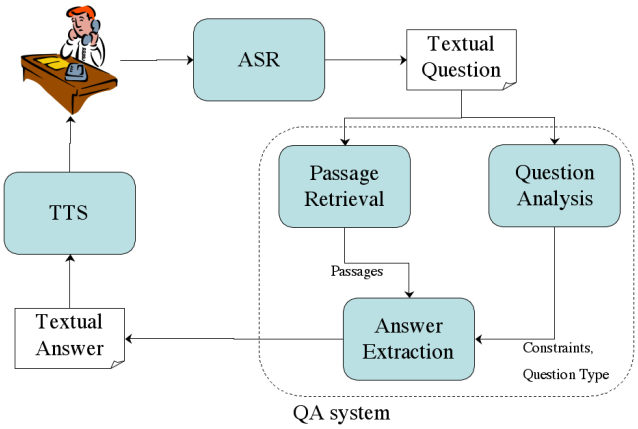

Figure 1: Modules of a voice-activated Question Answering system

It is particularly important to reduce the amount of recognition errors as much as possible, as they can imply strong

modifications in the meaning of the original sentence. Actually, these errors become crucial in the case of Named Entities (NEs), since they are usually some of the most meaningful words in the question. For instance, in the question "What is the capital of Iran?", recognizing "Iran" as "Iraq" is an error far more important than an error on any other word. Unfortunately, NEs are often very difficult to be recognized properly (sometimes because they are in a language different to the user's one), so this fact represents one of the biggest open challenges in voice-activated QA. Moreover, if the QA system uses an n-gram based passage retrieval (PR) engine, all the ASR mistakes can have a negative effect on the search, as they may lead to retrieve n-grams that were not included in the original sentence. NEs can be characterized by their high IDF (Inverse Document Frequency) and their redundancy in the retrieved passages. Our hypothesis is that recognition errors on question words which have a high IDF and are redundant in the set of retrieved passages are key, independently if they occur on NEs or not.

The aim of this work is to study the correlation between the recognition errors on question words with the above characteristics and the resulting errors in the Passage Retrieval module. We chose to limit our study to this phase and not to the full QA system because the errors in the question analysis and answer extraction phases are so important that they can mask the retrieval errors as noted by [3]. We computed the IDF of the words of the original sentence that were misrecognized in the ASR process both over the document collection and the passages retrieved by the PR engine using the full correct sentence. We carried out this experiment for several language models with a different number of Named Entities in each of them.

The rest of the paper is structured as follows. In Section 2 we present the Passage Retrieval system used for this study. In Section 3 a brief discussion about some interpretations of the IDF weight is provided. Then in Section 4 we explain the experiments we have performed and present and discuss the obtained results. Finally, we draw some conclusions.

2. THE X PASSAGE RETRIEVAL SYSTEM

In our study, we have used the X[1] Passage Retrieval system. This PR system uses a weighting scheme based on n-grams density. This approach has been proved to be more effective in the PR and QA tasks than other commonly used IR systems based on keywords and the well-known TF.IDF weighting scheme (citation omitted). So, X works under the premise that, in a sufficiently large document collection, question n-grams should appear near the answer at least once. The architecture of X is shown in Figure 2.

The first step consists in extracting passages which contain question terms from the document collection, which is done using the standard TF.IDF scheme. Subsequently, the system extracts all question k-grams (with $1 \leq k \leq n$, where n is the number of terms of the question) from both the question and each of the retrieved passages. The output of the system is a list of at most M passages (in our experiments we set $M = 30$) re-ranked according to a similarity value calculated between the passages and the question. The similarity between the question q and a passage p is defined in Equation 1.

[1]anonymized for blind review

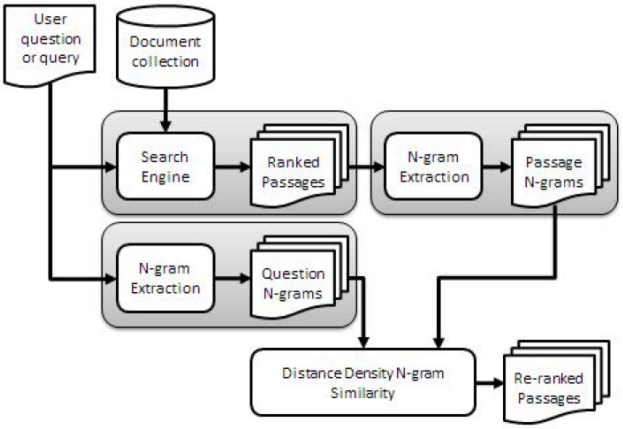

Figure 2: Structure of the X Passage Retrieval engine

$$Sim(p,q) = \frac{\sum_{\forall x \in (P \cap Q)} \frac{h(x)}{1 + \alpha \cdot ln(1 + d(x, x_{max}))}}{\sum_{i=1}^{n} w(t_{q_i})} \quad (1)$$

In this equation P is the set of k-grams ($1 \leq k \leq n$) contained in passage p and Q is the set of k-grams in question $q = (t_{q_1}, \ldots, t_{q_n})$; n is the total number of terms in the question. $w(t)$ is the term-weight, determined by:

$$w(t) = 1 - \frac{log(n_t)}{1 + log(N)} \quad (2)$$

Here n_t represents the number of sentences in which the term t occurs and N is the number of sentences in the collection.

The weight of each k-gram $x = (t_{x_1}, \ldots, t_{x_k})$ is calculated by means of the function $h(x) = \sum_{j=1}^{k} w(t_{x_j})$.

Finally, the distance $d(x, x_{max})$ is calculated as the number of words between any k-gram x and the one having maximum weight (x_{max}). α is a factor, empirically set to 0.1, that determines the importance of the distance in the similarity calculation.

3. ESTIMATING THE INFORMATIVENESS OF A TERM

It is a matter of fact that, given a collection of documents, it is very likely that each word will appear in a different number of documents. The two extreme cases are that a word appears in all the documents of the collection or that it is found in just one of them. In the first case, we can intuitively say that the word will not be very informative for none of those documents, since it makes no distinction between them. However, in the other case it is clear that the word is very probably one of the most informative for

that passage, as it is the only one that contains it. This idea, extended to all the range between the two presented cases, is the one that underlies the IDF formula [2]. Thus, the IDF weight for a word w in a document collection D can be written as

$$IDF = -log(\frac{|D(w)|}{|D|}) \quad (3)$$

Where $|D(w)|$ is the number of documents where the word w is contained and $|D|$ is the number of documents in the collection. An IDF weight equal to zero indicates that the word appears in all the documents of the collection and the higher this value is, the more relevant the word is in the collection.

Furthermore, from the point of view of Passage Retrieval, two interpretations can be given to the IDF weight, depending on the set of documents over which the formula is applied. On one hand, if we take the complete document collection and calculate the IDF over it this value gives an idea of how important is the document in the collection, which is called *term informativeness*. On the other, if this set of documents is reduced to those that the PR retrieval engine returned given a query in which the word appeared, the IDF computed for this word and these documents can be interpreted as the redundancy of the term. In other words, this measure gives an idea of how important has the word been for the PR engine in the process of searching the relevant passages.

4. EXPERIMENTS AND RESULTS

For our experiments we have used the questions in Spanish language from the CLEF[2] QA 2003-2006 contests. The target collection (the set of documents to be searched in order to find the answer) is composed by documents of the EFE (Spanish news agency) of the years 1994 and 1995. The set of questions amounts to $1,800$ questions divided in two subsets: $1,600$ for training and 200 for test. The 200 test questions were acquired by an specific user and are used as the input of the ASR.

For our experimentation, we have trained a generic language model for the ASR with just the training questions separating the NEs in a category. Then, in order to carry out different experiments, we have added more elements to this set according to its frequency in the document collection. So, we can distinguish two types of language models, namely, the Open Named Entity models, which include only the N most frequent NEs taken from the target collection and the Closed Named Entity models, which include all the test NEs and the N most frequent NEs taken from the same collection. In both cases the minimum number of NEs considered in the category was $4,000$ and the maximum $48,000$. As the original corpus does not have the NEs tagged in any way, previously to this process we automatically tagged the corpus using a POS-tagger.

Once all the test questions have been recognized using one of these models, we have considered two outputs: the recognized sentences themselves and the Word Error Rate (WER). Then, we have performed the Passage Retrieval process, taking the recognized sentences as its input.

The output of the Passage Retrieval phase is a ranked list of passages. So, it is interesting to know if this proposed ranking would match what a user would expect from the PR system. In the IR task there are some measures that are commonly used to take into account the position of the retrieved relevant results. Among the available measures, we selected the Normalized Discounted Cumulative Gain (nDCG) since it is the measure that best models user preferences, according to [5].

In order to calculate IR measures such as nDCG, it is necessary to have a set of *relevance judgments*, which is a set of documents considered to be relevant for the query. In our case, the set was built using hand-made answer patterns and regular expressions used to detect whether a passage contained the answer or not.

Normalized DCG at position π is defined as:

$$nDCG_\pi = \frac{DCG_\pi}{IDCG_\pi} \quad (4)$$

where $IDCG_\pi$ is the "ideal" DCG obtained by ranking all relevant documents at the top of the ranking list, in order of relevance, and $DCG_\pi = rel_1 + \sum_{i=2}^{\pi} \frac{rel_i}{\log_2 i}$, where rel_i is the degree of relevance of the result at position i.

As exposed in Section 2, X is an n-gram based PR engine. So, we can see the ASR process that works before the PR as a "noise introducer", in the sense that it can modify the original sentence n-grams. Thus, it would be interesting to relate the nDCG values obtained for each of the language models to that which would be obtained if the PR process was performed taking as its input the correct test questions. For this reason, we have used as the measure of the Passage Retrieval performance for each language model the value $nDCG(ref\_sents) - nDCG(recognized\_sents)$ (in the following tables we will refer to this as *nDCG diff*). The nDCG obtained for the original test set (with no ASR errors) is 0.584 (average over the set of 200 questions).

Finally, we have also calculated the term informativeness and the redundancy of the words of the original sentences that were misrecognized in the ASR process. These calculations were done over the complete target collection and the passages retrieved by the PR engine using the full correct sentence (not the recognized sentence that can have errors). We have calculated a composition of the misrecognized words of each sentence, both using the mean and max operators and, for each language model, we have averaged the results obtained for each sentence.

In order to avoid zeroes when the word does not appear in any of the documents returned by the PR engine, in the case of the redundancy we have slightly modified the IDF formula adding one to both elements of the fraction as shown in equation 5.

[2]http://www.clef-campaign.org

$$redundancy = -log(\frac{|D(w)|+1}{|D|+1}) \quad (5)$$

The obtained results are presented in tables 1 and 2.

Table 1: Closed Entity Model results

		avg redund.			
# NE	WER	mean	max	Term inf	nDCG diff
4000	0.265	0.348	0.529	2.329	0.151
8000	0.298	0.419	0.606	2.020	0.183
12000	0.305	0.432	0.614	2.011	0.197
16000	0.310	0.448	0.636	2.102	0.192
20000	0.310	0.454	0.644	2.143	0.192
24000	0.306	0.456	0.648	2.091	0.195
28000	0.312	0.461	0.660	2.159	0.201
32000	0.319	0.487	0.689	2.241	0.208
36000	0.319	0.489	0.691	2.293	0.205
40000	0.319	0.496	0.698	2.330	0.199
44000	0.321	0.493	0.698	2.298	0.203
48000	0.321	0.493	0.698	2.298	0.204

Table 2: Open Entity Model results

		avg redund.			
# NE	WER	mean	max	Term inf	nDCG diff
4000	0.333	0.522	0.755	2.379	0.265
8000	0.347	0.530	0.762	2.526	0.262
12000	0.351	0.531	0.750	2.455	0.273
16000	0.350	0.533	0.759	2.364	0.252
20000	0.348	0.534	0.760	2.389	0.248
24000	0.342	0.531	0.760	2.292	0.246
28000	0.344	0.526	0.755	2.297	0.242
32000	0.342	0.533	0.764	2.336	0.232
36000	0.344	0.534	0.766	2.388	0.229
40000	0.342	0.539	0.768	2.409	0.222
44000	0.345	0.536	0.768	2.390	0.226
48000	0.345	0.535	0.768	2.375	0.226

The difference in behaviour with respect to the growing number of NEs is due to the "open" vs. "closed" nature of the models: in the closed entity model, the fewer are the number of NEs in the model, the lesser are the probabilities of committing an error. Increasing the number of NEs leads to a higher probability of committing an error by recognizing a NE for another one. This behaviour is opposed to what happens with the open entity model, where the introduction of new NEs increases the chances of recognizing the right one.

With regard to the relationship between redundancy and nDCG in the retrieved passages, it can be observed that in the closed entity model, the lower the redundancy of the misrecognized term, the lower is the error in nDCG (see Figure 3). In the open model (Figure 4) this correlation is not observed. In both models, no correlation has been found between the nDCG and the IDF of the misrecognized terms, somehow surprisingly as we expected that errors on terms with high IDF should be more important.

Our interpretation of these results is that in the closed entity model the errors on NEs are less frequent, therefore there are more errors on non-NE words, which are also words that are repeated frequently in the result set (i.e., redundant).

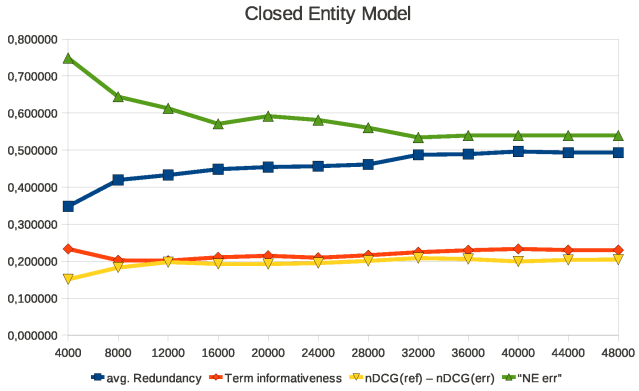

Figure 3: Closed Entity Model Results. Term informativeness values have been divided by 10. "En err": error on NEs.

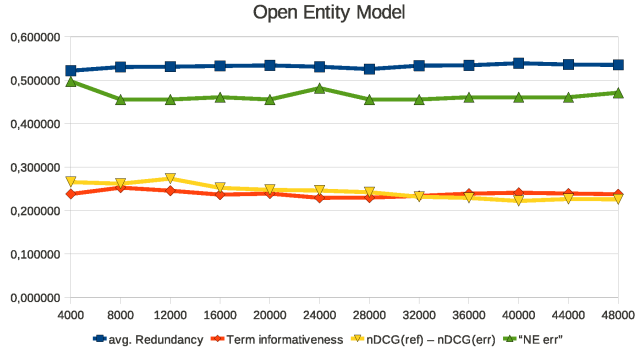

Figure 4: Open Entity Model results. Term informativeness values have been divided by 10. "En err": error on NEs.

As it can be observed from Figure 3, the error on NEs is inversely proportional to redundancy, indicating that NEs are less redundant than other kind of words.

5. CONCLUSIONS

In this paper we attempted to find a relationship between the redundancy and term informativeness of misrecognized terms on the output of a PR module of a voice-activated QA system. We used ASR using a closed and an open NE model. Our results show that term informativeness, measured as IDF, is not an indicator of whether the error on that term will be relevant or not for the passage retrieval process. On the other hand, the redundancy of a term in the retrieved passages seems to be an important clue on whether an error on that term will produce a worse result, at least if a closed NE model is used.

6. REFERENCES

[1] S. Harabagiu, D. Moldovan, and J. Picone. Open-domain voice-activated question answering. In *Proceedings of the 19th international conference on Computational linguistics*, COLING '02, pages 1–7, 2002.

[2] K. Jones. Index term weighting. *Information Storage and Retrieval*, 9(11):619–633, 1973.

[3] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu. Performance Issues and Error Analysis in an Open-Domain Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 133–154, New York, USA, 2003.

[4] P. Rosso, L.-F. Hurtado, E. Segarra, and E. Sanchis. On the voice-activated question answering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, PP(99):1 –11, 2010.

[5] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM.

[6] J. Turmo, P. Comas, S. Rosset, O. Galibert, N. Moreau, D. Mostefa, P. Rosso, and D. Buscaldi. Overview of QAST 2009. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece*, volume 6241 of *Lecture Notes in Computer Science*, pages 197–211. Springer, 2009.